软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(12):3846−3861 [doi: 10.13328/j.cnki.jos.005606]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

联合 Laplacian 正则项和特征自适应的数据聚类算法*

郑建炜[1], 李卓蓉[1,2], 王万良[1], 陈婉君[1]

[1](浙江工业大学 计算机科学与技术学院,浙江 杭州 310023)

[2](浙江大学城市学院 计算机与计算科学学院,浙江 杭州 310015)

通讯作者: 王万良, E-mail: wwl@zjut.edu.cn

摘 要: 在信息爆炸时代,大数据处理已成为当前国内外热点研究方向之一.谱分析型算法因其特有的性能而获得了广泛的应用,然而受维数灾难影响,主流的谱分析法对高维数据的处理仍是一个极具挑战的问题.提出一种兼顾维数特征优选和图 Laplacian 约束的聚类模型,即联合拉普拉斯正则项和自适应特征学习(joint Laplacian regularization and adaptive feature learning,简称LRAFL)的数据聚类算法.基于自适应近邻进行图拉普拉斯学习,并将低维嵌入、特征选择和子空间聚类纳入同一框架,替换传统谱聚类算法先图 Laplacian 构建、后谱分析求解的两级操作.通过添加非负加和约束以及低秩约束,LRAFL 能获得稀疏的特征权值向量并具有块对角结构的 Laplacian 矩阵.此外,提出一种有效的求解方法用于模型参数优化,并对算法的收敛性、复杂度以及平衡参数设定进行了理论分析.在合成数据和多个公开数据集上的实验结果表明,LRAFL 在效果效率及实现便捷性等指标上均优于现有的其他数据聚类算法.

关键词: Laplacian 矩阵;特征选择;谱聚类;相似度矩阵;低秩约束

中图法分类号: TP391

中文引用格式: 郑建炜,李卓蓉,王万良,陈婉君.联合 Laplacian 正则项和特征自适应的数据聚类算法.软件学报,2019,30(12):3846−3861. http://www.jos.org.cn/1000-9825/5606.htm

英文引用格式: Zheng JW, Li ZR, Wang WL, Chen WJ. Clustering with joint Laplacian regularization and adaptive feature learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(12):3846−3861 (in Chinese). http://www.jos.org.cn/1000-9825/5606.htm

Clustering with Joint Laplacian Regularization and Adaptive Feature Learning

ZHENG Jian-Wei[1], LI Zhuo-Rong[1,2], WANG Wan-Liang[1], CHEN Wan-Jun[1]

[1](School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

[2](School of Computer and Computing Science, Zhejiang University City College, Hangzhou 310015, China)

Abstract: The explosion of information has been evoking a leading wave of big data research during recent years. Despite many empirical successes of spectral clustering algorithms, it is still challenging to cluster the high dimensional data due to the curse of dimensionality. This study proposes a novel algorithm referred to as joint Laplacian regularization and adaptive feature learning (LRAFL), which adaptively learns the feature weights and fits the feature selection as well as clustering into a unified framework, rather than the two-phase strategy of typical approaches. With a new rank constraint imposed on the Laplacian matrix, the connected components in the resulted similarity matrix are exactly equal to the cluster number. An effective approach is also proposed to solve the formulated optimization problem. Comprehensive analyses, including convergence behavior, computational complexity, and together with parameter determination are also presented. Surprisingly sound experimental results can be achieved on synthetic data and benchmark datasets by the proposed algorithm when compared with the related state-of-the-art clustering approaches.

* 基金项目: 国家自然科学基金(61602413, 61873240); 浙江省自然科学基金(LY19F030016)

Foundation item: National Natural Science Foundation of China (61602413, 61873240); Natural Science Foundation of Zhejiang Province of China (LY19F030016)

收稿时间: 2016-12-03; 修改时间: 2017-06-09, 2017-12-07; 采用时间: 2018-05-17

Key words: Laplacian matrix; feature selection; spectral clustering; similarity matrix; low-rank constraint

聚类分析是数据挖掘领域的研究热点之一,旨在无标签情形下对数据进行分组,使组内数据尽可能相似而组间数据尽可能不同,被广泛应用于图像分割[1]、目标分簇[2]、深度学习模型[3]等科学应用领域.在大数据背景下,实际输入数据除“海量样本”特点外,还具有极高的特征维数.以在线文本数据为例,当采用矢量空间描述每个文档时,大词汇量往往导致样本维数达到 5 000 以上.此外,一张解析度为 256×256 的图像矢量化后的维数则是 65 536.受“维数灾难”限制,对高维数据进行合理高效的聚类分析是一个极具挑战性的问题.过高的样本维度包含冗余的特征信息和异常噪声,不仅降低了后续聚类操作的运算效率,也影响了其他性能指标.针对该问题,常见的思路是引入特征选择进行维数预约简,然后在子空间进行相似度矩阵构建并对嵌入数据实施谱聚类分析.

特征选择(feature selection,简称 FS)[4]从原始样本空间中挑选最具代表性的维数子集,其核心问题依据特定准则评价各子集的优劣并确定选择结果.传统搜索策略[5]的缺点是直接利用数据的统计指标对每个特征进行单独评分并取分值较高者为结果集,缺乏整体的优劣评判标准[6].针对此问题,学者们开展了联合特征选择研究,通过稀疏正则化约束[7]进行特征选择并兼顾子空间学习.Cai 等人[8]结合流形学习和 l_1 正则化模型进行稀疏的联合特征选择.所选用的 l_1 范数虽然意义明确,但其稀疏性仅作用于独立的特征点[9].更多的算法[10]通过对投影矩阵约束 $l_{2,1}$ 范数以保证行稀疏,选择矩阵非零行对应的特征集合为最优特征子集.在评价准则方面,通常选择能有效保持数据本质结构的特征,采用图论模型刻画全局结构、局部流形以及鉴别性信息等.多簇特征选择法(multi-clusters FS,简称 MCFS)[8]首先计算高维数据的低维流形嵌入,然后对投影矩阵采用 l_1 范数进行稀疏约束,根据回归系数对每个特征进行排序,最终选择最易保持局部流形结构的特征.局部学习聚类(local learning based clustering for FS,简称 LLCFS)[11]将特征关联性引至内置的正则化局部学习模型,使得演化的 Laplacian 图能够迭代优化.自适应结构学习(FS with adaptive structure learning,简称 FSASL)[12]旨在结合全局信息挖掘以及局部流形学习进行样本结构保持,兼顾了稀疏性和保局性两种优势.局部保持得分法(locality preserving score,简称 LPS)[13]则从误差抑制的角度出发,对每个特征的重构能力进行排序,获得最优的子特征集.上述算法采用独立的步骤按序进行子空间学习和聚类操作,其弊端是无法达到聚类目标的整体最优效果.常见的解决方案是将子空间聚类融合为联合优化整体,通过聚类指标和降维指标互相反馈优化模型的各约束项.鉴别嵌入聚类法(discriminative embedded clustering,简称 DEC)[14]联合 Fisher 鉴别投影和 *k*-means 提出一致性的分簇框架,但其受限于 *k*-means 的本质约束,无法适应单流形多环分布数据.非负鉴别法(nonnegative discriminative FS,简称 NDFS)[15]将聚类标签反馈于特征选择步骤,提升了特征子集的鉴别性,然而其特征选择过程缺乏结构性意义,且算法容易陷入局部最优点.

提升相似度矩阵(或称为关联矩阵、邻接矩阵)结构是进一步改进子空间聚类性能的关键思想,也是谱聚类算法的核心步骤.Wang 等人[16]基于局部线性嵌入思想[17]构建 Laplacian 图,获得了良好的标签传播性能.Elhamifar 等人[18]则以全局线性表示系数作为关联矩阵构建基础,通过 l_1 范数提出了稀疏子空间聚类法(sparse subspace clustering,简称 SSC).SSC 假设每个数据由同一子空间中其他样本稀疏表示,挖掘不同组的表示关系,但其缺乏空间分布结构考虑.Liu 等人[19]提出了低秩表示(low-rank representation,简称 LRR)聚类法,利用核范数约束系数矩阵,获得更好的全局性.SSC 和 LRR 以输入样本子空间相互独立或正交为假设,其理想状态下的相似矩阵具有刻画子空间属性的块对角结构.进一步,Lu 等人[20]给出了一组强制块对角条件,并指出:在数据充分并且子空间相互独立的前提下,正则项满足该条件可保证相似矩阵具有块对角结构.Feng 等人[21]将对应的拉普拉斯矩阵进行低秩约束,并添加至 SSC 和 LRR 以保证块对角状态,获得更优的相似度结构.此外,在系数矩阵优化问题上,新晋算法都采用 Laplacian 正则项约束提升相似度矩阵的块对角结构[22−24],非负稀疏 Laplacian 正则约束的 LRR 模型(non-negative sparse Laplacian regularized LRR,简称 NSLLRR)[22]以非负性、稀疏性为条件,增加超图拉普拉斯约束,具有良好的样本表示能力.Hu 等人[23]提出的光滑表示聚类模型(smooth representation clustering,简称 SMR)基于增强型组效应条件进行相似性度量,算法在保证高质量聚类性能的前提下获得了大幅度效率提升.为更好地逼近低秩结构,分组低秩结构模型(low-rank structure,简称 LRS)[24]引入组指示规化对各

簇样本进行 Schatten p 范数正则项约束,其缺陷是抗噪性差且模型运算效率较低.

综上所述,现存的特征选择型算法缺乏样本间关联结构描述,导致次优的聚类性能;而 Laplacian 正则型表示模型则都采用原始数据直接构建关联矩阵,独立于表示系数更新操作,也不具备整体算法的最优性.虽然鲁棒子空间分割法(robust subspace segmentation,简称 RSS)[25]实现了重构系数和相似度矩阵的兼顾学习,具有更优的结构挖掘能力,但缺乏特征优选机制,对现实高维数据的抗噪性弱,且其自表示框架受稀疏性、非负性等约束的影响,运行效率较低,样本规模尺度化能力有待进一步提高.针对现存算法的问题,本文基于自适应近邻进行图拉普拉斯学习,将低维嵌入、特征选择和簇结构学习纳入同一框架,提出一种兼顾自适应特征优选和簇结构学习的聚类模型,即联合拉普拉斯正则项和自适应特征学习(joint Laplacian regularization and adaptive feature learning,简称 LRAFL)的数据聚类算法,具体工作如下.

1) 提出一种图Laplacian矩阵更新策略,保证其秩结构与目标聚类数的一致性,使得模型优化结果直接具备分簇块对角结构,规避了后续 k-means、谱分解等操作;
2) 将特征学习机制融入Laplacian矩阵构建框架,在保证噪声特征抑制的前提下,去除高复杂度的表示系数学习过程,提升模型求解效率;
3) 设计具备唯一最优解的参数优化方案,对模型部分待定参数进行推演分析,给出更具指示意义的设定方法,进一步加速模型实现效率.

1 相关工作

本节介绍谱聚类算法中的两个关键步骤,即相似度矩阵构建和拉普拉斯正则约束项,其中,前者用于挖掘数据分布结构,而后者是引导块对角状态的核心技术.

1.1 相似度矩阵构建

传统的相似图构建方法如ε邻域图、k 近邻图、全连接图等都存在着明显的缺陷,包括:(1) 分析尺度选择困难;(2) 参数敏感性强;(3) 多尺度数据适应度弱;(4) 抗噪性差等等.为解决存在的问题,Wang 等人[16]在邻域图基础上通过线性表示计算相似度权值,提升了算法抗噪性;Zelnik-Manor 等人[26]提出自校正谱分簇算法,缓解了第 1 个和第 3 个缺陷.Cheng 等人[27]引入稀疏表示进行邻域图构建,可以有效解决第 1 个和第 4 个问题,也规避了高敏感性的待定参数ε和 k,但其正则项参数的敏感性仍然较强,且存在 l_1 范数求解运行效率低的问题.Huang 等人[28]采用非负和加权限制替代文献[27]中的 l_1 范数约束,提出了单纯型稀疏表示(simplex sparse representation,简称 SSR)邻域图构建方法,能有效解决上述前 3 个问题;而且算法不需要人工设定参数,运行效率和实现简易度亦优于其他对比算法.

给定数据集 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]\in R^{m\times n}$,其中,$\boldsymbol{x}_i$ 是第 i 个 m 维输入样本,n 是训练样本总数.定义邻域图模型 $\boldsymbol{S}$,其元素 s_{ij} 表示数据点 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 互为近邻的概率,$\boldsymbol{s}_i\in R^n$ 表示 $\boldsymbol{S}$ 的第 i 个列向量.SSR 的目标函数为

$$\left.\begin{aligned}&\min_{\boldsymbol{s}_i}\|\boldsymbol{X}_{-i}\boldsymbol{s}_i-\boldsymbol{x}_i\|_2^2\\&\text{s.t. }\boldsymbol{s}_i\geqslant 0,\boldsymbol{s}_i^T\boldsymbol{1}=1\end{aligned}\right\}\tag{1}$$

其中,$\boldsymbol{X}_{-i}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{i-1},\boldsymbol{0},\boldsymbol{x}_{i+1},\ldots,\boldsymbol{x}_n]$表示剔除第 i 个输入数据的训练样本集,$\boldsymbol{0}$ 是 $m\times 1$ 的零向量,$\boldsymbol{1}$ 是元素全为 1 的 $n\times 1$ 向量.公式(1)通过重构表示能力说明高权值系数的成对样本具有更高的概率互为近邻,具有天然的样本稀疏性和奇异点抗噪性,其缺点是不具备特征稀疏性,因此不适于高维度冗余数据应用.

1.2 Laplacian正则约束

Laplacian 矩阵构建的方式多样,且各算法的作者都称自己的谱分析矩阵为 Laplacian.给定对称的相似度矩阵 $\boldsymbol{S}$,RatioCut[29]所构建的 Laplacian 矩阵为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$,其中,对角矩阵 $\boldsymbol{D}$ 称为度矩阵,相应的对角元素 $d_i=\sum_{j=1}^n s_{ij}$. NCut[30]将上述 $\boldsymbol{L}$ 进行规范化操作,即 $\boldsymbol{L}_s=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$ 或 $\boldsymbol{L}_{ns}=\boldsymbol{D}^{-1}\boldsymbol{L}$,其中,前者是对称矩阵而后者是非对称矩阵.当给定非对称的 $\boldsymbol{S}$ 时,则相应的非规范化 Laplacian 矩阵计算为 $\boldsymbol{L}=\boldsymbol{D}-(\boldsymbol{S}^T+\boldsymbol{S})/2$[13,31],其中,度矩阵 $\boldsymbol{D}$ 的对角元素为

$d_i=\sum_j(s_{ij}+s_{ji})/2$. 在经典谱聚类算法中,无论 Laplacian 矩阵形式如何,后续操作都对该矩阵进行特征求解,并针对前 c 个特征矢量进行 k-means 聚类,其中,c 是数据簇结构目标.

最近,Hu 等人[23]以谱聚类为基础,结合低秩重构表示思想将一般的表示型谱聚类模型归纳为

$$\min_{\boldsymbol{Z}}\alpha\|\boldsymbol{X}-A(\boldsymbol{X})\boldsymbol{Z}\|_l+\Omega(\boldsymbol{X},\boldsymbol{Z}) \tag{2}$$

其中,$\alpha>0$ 是平衡参数,$A(\boldsymbol{X})$表示字典矩阵,$\boldsymbol{Z}$ 是系数矩阵,$\|\cdot\|_l$ 表示合适的范数.公式(2)前半部分刻画了重构表示 $A(\boldsymbol{X})\boldsymbol{Z}$ 逼近数据 $\boldsymbol{X}$ 的程度,后半部分是 Laplacian 谱约束正则项.

稀疏子空间聚类对公式(2)的正则项采用某种稀疏度量,从而使 $\boldsymbol{Z}$ 具有特定目标结构.常见的包括 SSC 的 l_1 范数约束、LRR 的核范数约束以及 SMR 和 RSS 的组效应约束等等.考虑到块对角结构的相似矩阵能更好地刻画簇结构属性,Feng 等人[25]利用相似度矩阵对角块个数与 Laplacian 矩阵秩约束之间的关系,对图拉普拉斯矩阵添加秩约束:

$$K=\left\{\boldsymbol{Z}\mid rank(\boldsymbol{L}_S)=n-c,\boldsymbol{W}=\frac{1}{2}(|\boldsymbol{Z}|+|\boldsymbol{Z}^T|)\right\} \tag{3}$$

其中,c 表示对角块的个数,也即簇目标数.将上述秩约束添加至子空间聚类模型,可保证清晰的块对角结构,具体目标模型描述为

$$\left.\begin{aligned}&\min_{\boldsymbol{Z}}\left(\|\boldsymbol{Z}\|_1+\frac{\lambda}{2}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_F^2\right)\\&\text{s.t. }diag(\boldsymbol{Z})=0,\boldsymbol{Z}\in K\end{aligned}\right\} \tag{4}$$

其中,λ是平衡参数;$diag(\boldsymbol{Z})=0$ 用于约束对角元素 $z_{ii}=0$,以避免平凡解.

2 LRAFL 算法描述

结合现有工作,考虑到自适应邻域学习和块对角 Laplacian 矩阵对聚类效果的重要性以及自表示学习的复杂性,本节将公式(1)和公式(4)中的表示系数转变为邻域结构约束,并辅以稀疏性、参数自学习、特征寻优以及簇结构直接确定等优势,提出一种兼顾特征选择和谱聚类的算法 LRAFL.首先对该算法目标函数的构建过程进行描述,然后给出了模型求解优化方案.

2.1 目标函数构建

探索数据的局部连通性,即相似度权值,是聚类任务的典型策略[32].根据本文开始部分的描述,常规的表示系数[18,19]和线性关联[16]都存在计算效率低以及缺乏全局最优等弊端,本节直接以相似度计算为基础,辅以特征加权、低秩块对角约束等构建目标函数.首先给定任意输入数据 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$,其距离 $\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2$ 与相似度权值 s_{ij} 应呈反比关系,即短距离对应大权值、长距离对应小权值.因此,结合公式(1)对权值的概率条件约束,一种自然的相似度计算方法为

$$\min_{\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0}\sum_{j=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\,s_{ij} \tag{5}$$

然而,公式(5)具有平凡解,仅 $\boldsymbol{x}_i$ 的最近邻样本获得概率相似度 1 而余下的 $s_{ij}=0$.另一方面,如果在不包含任何距离信息约束下求解式:

$$\min_{\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0}\sum_{j=1}^{n}s_{ij}^2 \tag{6}$$

则得到另一种平凡解,即所有样本都是 $\boldsymbol{x}_i$ 的近邻且概率相似度为 $1/n$,可以看作相似度赋值的先验分布,其本质则是 l_2 范数约束条件[33].结合公式(5)和公式(6),$\boldsymbol{x}_i$ 的邻域相似度计算为

$$\min_{\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0}\sum_{j=1}^{n}\left(\frac{1}{2}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\,s_{ij}+\beta s_{ij}^2\right) \tag{7}$$

其中,第 2 项为正则项,β是正则化参数.联合所有的输入数据 $\boldsymbol{x}_i,i=1,\ldots,n$,则完整的相似度计算可以描述为

$$\min_{\forall i,\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0}\sum_{i,j=1}^{n}\left(\frac{1}{2}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2 s_{ij}+\beta s_{ij}^2\right) \tag{8}$$

通过第2.2节模型优化求解过程可知,公式(8)中各相似度矢量$\boldsymbol{s}_i$具有稀疏的闭式解,模型优化效率高且能够有效抑制奇异噪声样本.

其次,为引入特征优选机制,使算法具有奇异特征抑制性能,采用特征加权因子$\boldsymbol{w}\in R^{m\times 1}$将公式(8)调整为

$$\min_{\forall i,\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0}\sum_{i,j=1}^{n}\left(\frac{1}{2}\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2 s_{ij}+\gamma s_{ij}^2\right) \tag{9}$$

其中,⊙表示元素相乘符号.与公式(5)和公式(6)类似,直接以公式(9)为目标函数会出现平凡解.即:当$\boldsymbol{w}$取零向量且相似度为1/n时,模型值最小.因此,进一步将相似度约束条件添加至$\boldsymbol{w}$权值矢量,即:

$$\left.\begin{aligned}&\min_{\boldsymbol{s},\boldsymbol{w}}\sum_{i,j=1}^{n}\left(\frac{1}{2}\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2 s_{ij}+\beta s_{ij}^2\right)+\gamma\|\boldsymbol{w}\|_2^2\\&\text{s.t. }\boldsymbol{w}^T\mathbf{1}=d,w_i\geqslant 0,\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0\end{aligned}\right\} \tag{10}$$

其中,$d\leqslant m$表示选择后有效特征数.公式(10)第1部分用于相似度矩阵构建,子项$\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2 s_{ij}$在特征优选约束下,使邻近的样本对具有更高的相似度权值,而非近邻样本对具有较低的相似度权值,余下部分是特征加权矢量和相似度值的l_2范数约束,用于规避平凡解并引导模型未知量具有光滑的数值结构.

文献[14]等聚类算法通过投影矩阵进行特征提取,相比较而言,公式(10)采用特征选择操作拥有的优势包括:(1) 所采用的矢量操作较特征提取算法的特征分解操作效率更高;(2) 对于输入数据不同特征的支撑作用具有更加明确的物理意义;(3) 可以在不指定特征子集规模的前提下进行加权赋值,而特征提取必须指定子空间维数.此外,通过对公式(10)模型中相似矩阵$\boldsymbol{S}$和特征权值矢量$\boldsymbol{w}$进行交替优化,可同时实现流形结构学习和联合特征选择.通过$\boldsymbol{S}$的迭代更新和优化,使得近邻关系具有自适应性,从而确保特征选择及谱聚类不再基于固定不变的图Laplacian结构.

与其他谱聚类算法相似,公式(10)得到的相似矩阵$\boldsymbol{S}$不能直接用于数据聚类,需进行谱分析且利用k-means得到聚类结果[32].根据定理1可知:当Laplacian矩阵$\boldsymbol{L}_S$的秩为$n–c$时,则相应的相似矩阵$\boldsymbol{S}$恰好具有c分簇对角结构,无需额外的k-means操作.为实现该目标,将文献[20]中的低秩约束(即公式(3))引入公式(10),则有:

$$\left.\begin{aligned}&\min_{\boldsymbol{s},\boldsymbol{w}}\sum_{i,j=1}^{n}\left(\frac{1}{2}\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2 s_{ij}+\beta s_{ij}^2\right)+\gamma\|\boldsymbol{w}\|_2^2\\&\text{s.t. }\boldsymbol{w}^T\mathbf{1}=d,w_i\geqslant 0,\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0,rank(\boldsymbol{L}_S)=n-c\end{aligned}\right\} \tag{11}$$

其中,$rank(\boldsymbol{L}_S)=n–c$约束项与定理1中的零特征值重根数等价.然而,直接对公式(11)求解非常困难[21],本文根据命题1进一步将公式(11)调整为

$$\left.\begin{aligned}&\min_{\boldsymbol{s},\boldsymbol{w}}\sum_{i,j=1}^{n}\left(\frac{1}{2}\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2 s_{ij}+\beta s_{ij}^2\right)+\gamma\|\boldsymbol{w}\|_2^2+\lambda tr(\boldsymbol{F}^T\boldsymbol{L}_s\boldsymbol{F})\\&\text{s.t. }\boldsymbol{w}^T\mathbf{1}=d,w_i\geqslant 0,\boldsymbol{s}_i^T\mathbf{1}=1,s_{ij}\geqslant 0\end{aligned}\right\} \tag{12}$$

其中,符号tr是矩阵的迹,$\boldsymbol{F}\in R^{n\times c}$是Laplacian矩阵$\boldsymbol{L}_s$相应$c$个最小特征值的特征矢量.公式(12)是最后的LRAFL模型目标函数,基于自适应邻域学习构建图Laplacian矩阵,将低维嵌入、特征选择和谱聚类纳入同一框架,并添加非负加和约束以及等价低秩约束,模型结果具有明确的块对角结构.

定理 1[21,32]. 相似矩阵$\boldsymbol{S}$对应的拉普拉斯矩阵$\boldsymbol{L}_s$中,特征值为0的重根数与相似矩阵$\boldsymbol{S}$中块结构的数量相等.

命题 1. 最小化$Tr(\boldsymbol{F}^T\boldsymbol{L}_s\boldsymbol{F})$与$rank(\boldsymbol{L}_s)=n–c$具有等价性,其中,$\boldsymbol{F}\in R^{n\times c}$.

证明:假设$\sigma_i(\boldsymbol{L}_s)$是Laplacian矩阵第$i$小的特征值,根据拉普拉斯矩阵的半正定性[32],$\sigma_i(\boldsymbol{L}_s)\geqslant 0$成立,因此对$\boldsymbol{L}_s$秩约束为$n–c$等同于约束$\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s)=0$.再根据Ky Fan定理[34],即:

$$\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s)=\min_{\boldsymbol{F}\in R^{n\times c},\boldsymbol{F}^T\boldsymbol{F}=\boldsymbol{I}} Tr(\boldsymbol{F}^T\boldsymbol{L}_s\boldsymbol{F}),$$

则命题 1 得证. □

2.2 模型优化求解

在公式(12)中,相似矩阵 $\boldsymbol{S}$ 和特征权值向量 $\boldsymbol{w}$ 相互耦合,投影矩阵 $\boldsymbol{F}$ 的构建又依赖于相似矩阵和拉普拉斯矩阵,因此不能直接对其求取闭合解.本节采用交替优化的方法,依次对不同未知变量进行单变量优化,其中,每一次迭代都是一个凸优化过程.

首先,当固定相似矩阵 $\boldsymbol{S}$ 时,则 $\boldsymbol{F}$ 由 $\boldsymbol{L}_s$ 的前 c 个最小特征值所对应的特征向量构成,因此 $\boldsymbol{F}$ 也是固定矩阵.$\boldsymbol{L}_s$ 是一个实对称半正定矩阵,通过奇异值分解可得到 $\boldsymbol{L}_s=\boldsymbol{L}\boldsymbol{L}^T$.从而,目标函数(12)可以调整为

$$\left.\begin{aligned}\min\left\{\sum_{i,j=1}^{n}\left(\frac{1}{2}\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2\,s_{ij}\right)+\gamma\|\boldsymbol{w}\|_2^2\right\}&=\min\{tr(\boldsymbol{W}\boldsymbol{X}\boldsymbol{L}_s\boldsymbol{X}^T)+\gamma\boldsymbol{w}^T\boldsymbol{w}\}\\&=\min\{tr(\boldsymbol{W}\boldsymbol{Y}\boldsymbol{Y}^T)+\gamma\boldsymbol{w}^T\boldsymbol{w}\}\\&=\min\sum_{i=1}^{d}\left\{\left(w_i\sum_{j=1}^{n}y_{ij}^2\right)+\gamma w_i^2\right\}\\\text{s.t. }\ w_i\geqslant 0,\boldsymbol{w}^T\boldsymbol{1}=d\end{aligned}\right\}\tag{13}$$

其中,$\boldsymbol{W}$ 是以 $\boldsymbol{w}$ 为对角元素的对角矩阵,$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{L}$,而 y_{ij} 是 $\boldsymbol{Y}$ 矩阵对应的元素.公式(13)是一个典型的二次规划问题,常见的数值最优化技术包括内映射牛顿法、有效集算法等[35]都能够对之进行迭代优化获得特征权值矢量 $\boldsymbol{w}$.为进一步提升效率,本文提出一种闭式求解方案,将公式(13)进一步调整为

$$\left.\begin{aligned}&\min\sum_{i=1}^{d}\left\{\left(w_i\sum_{j=1}^{n}y_{ij}^2\right)+\gamma w_i^2\right\}=\min\sum_{i=1}^{d}\{(w_iz_i)+\gamma w_i^2\}=\min\left\|\boldsymbol{w}+\frac{\boldsymbol{z}}{2\gamma}\right\|_2^2\\&\text{s.t. }\ w_i\geqslant 0,\boldsymbol{w}^T\boldsymbol{1}=d\end{aligned}\right\}\tag{14}$$

其中,向量 $\boldsymbol{z}$ 的元素为 $z_i=\sum_{j=1}^{n}y_{ij}^2$.将约束条件添加成拉格朗日正则项:

$$L(\boldsymbol{w},\eta,\boldsymbol{\alpha})=\frac{1}{2}\left\|\boldsymbol{w}+\frac{\boldsymbol{z}}{2\gamma}\right\|_2^2-\eta(\boldsymbol{w}^T\boldsymbol{1}-d)-\boldsymbol{\alpha}^T\boldsymbol{w},$$

并对其微分方程置 0,可得到特征矢量 $\boldsymbol{w}$ 的初步解为 $\boldsymbol{w}=(-\boldsymbol{z}/2\gamma+\eta)_+$,符号$(\cdot)_+$表示元素非负,其中,$\eta$和$\boldsymbol{\alpha}$是拉格朗日系数.根据约束条件 $\boldsymbol{w}^T\boldsymbol{1}=d$,即 $\sum_{j=1}^{d}(-z_j/2\gamma+\eta)=d$,可得 $\eta=1+\sum_{j=1}^{d}(z_j/2\gamma d)$,则 $\boldsymbol{w}$ 的解为

$$\boldsymbol{w}=\left(\frac{2\gamma d-\boldsymbol{z}d+\sum_{j=1}^{d}z_j}{2\gamma d}\right)_+\tag{15}$$

其次,当 $\boldsymbol{F}$ 和 $\boldsymbol{w}$ 固定时,则公式(12)变成:

$$\left.\begin{aligned}&\min\sum_{i,j=1}^{n}\left(\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2\,s_{ij}+\lambda\sum_{i,j=1}^{n}\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2^2\,s_{ij}+\beta s_{ij}^2\right)\\&\text{s.t. }\ \boldsymbol{s}_i^T\boldsymbol{1}=1,s_{ij}\geqslant 0\end{aligned}\right\}\tag{16}$$

其中,$\boldsymbol{f}_i\in c\times 1$ 是 $\boldsymbol{F}$ 矩阵的第 i 行向量.与公式(13)类似,公式(16)也是一个二次规划问题,可通过内映射牛顿法、有效集算法等进行迭代求解获得最优相似矩阵 $\boldsymbol{S}$.为实现闭式求解方案,设定 $g_{ij}^x=\|\boldsymbol{W}(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2$,$g_{ij}^f=\|\boldsymbol{f}_i-\boldsymbol{f}_j)\|_2^2$,则公式(16)可以等价为向量形式:

$$\left.\begin{aligned}&\min\sum_{i,j=1}^{n}\left\{\|\boldsymbol{w}^{1/2}\odot(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2\,s_{ij}+\lambda\sum_{i,j=1}^{n}\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2^2\,s_{ij}+\beta s_{ij}^2\right\}=\min\sum_{i=1}^{n}\left\|\boldsymbol{s}_i+\frac{\boldsymbol{g}_i}{2\beta}\right\|_2^2\\&\text{s.t. }\ \boldsymbol{s}_i^T\boldsymbol{1}=1,s_{ij}\geqslant 0\end{aligned}\right\}\tag{17}$$

其中,向量 $\boldsymbol{g}_i$ 的元素为 $g_{ij}=g_{ij}^x+\lambda g_{ij}^f$.考虑到不同 $\boldsymbol{s}_i$ 之间的独立性,可对各个相似向量分开求解优化,相应的拉格朗日乘子式为

$$L(\boldsymbol{s}_i,\eta,\boldsymbol{\xi}_i)=\frac{1}{2}\left\|\boldsymbol{s}_i+\frac{\boldsymbol{g}_i}{2\beta_i}\right\|_2^2-\eta(\boldsymbol{s}_i^T\mathbf{1}-1)-\alpha_i^T\boldsymbol{s}_i \tag{18}$$

通过对公式(18)中的 $\boldsymbol{s}_i$ 求导,并令结果为 0,可得初始解:

$$\boldsymbol{s}_i=\left(-\frac{\boldsymbol{g}_i}{2\beta_i}+\eta\right)_+ \tag{19}$$

再根据约束条件 $\boldsymbol{s}_i^T\mathbf{1}=1$,有 $\eta=1/n+\left(\sum_{j=1}^n g_{ij}\middle/2n\beta_i\right)$,即 $\boldsymbol{s}_i$ 的求解公式为

$$\boldsymbol{s}_i=\left(\frac{2\beta_i-n\boldsymbol{g}_i+\sum_{j=1}^n g_{ij}}{2n\beta_i}\right)_+ \tag{20}$$

综上所述,完整的 LRAFL 如算法 1 描述.值得注意的是:在公式(12)目标函数下,如忽略算法 1 的迭代框架,即先令 s_{ij}=1,依公式(13)求解特征权值 $\boldsymbol{w}$;再固定 $\boldsymbol{w}$,联合优化 $\boldsymbol{S}$ 和 $\boldsymbol{F}$,可得到 LRAFL 模型的独立优化版(Ind),获得目标函数的快速解.然而该版本以模型次优性为代价,其实际应用性能弱于算法 1.为有效平衡模型的实施性能和运行效率,通过设置收敛条件(见第 3 节描述),可使模型在 I_m<15 次迭代内停止.

算法 1. LRAFL 描述.

输入:数据集 $\boldsymbol{X}$,聚类目标 c,迭代总数 I_m,平衡参数γ,β,λ,有效特征数 d;

输出:具有 c 分块对角结构的相似矩阵 $\boldsymbol{S}$,特征加权向量 $\boldsymbol{w}$.

1. 初始化特征加权向量 $\boldsymbol{w}^0$,设λ=0,通过公式(20)得到初始相似矩阵 $\boldsymbol{S}^0$,并计算投影矩阵 $\boldsymbol{F}^0$;
2. 设迭代次数 t=1;
3. 固定相似矩阵和投影矩阵,依公式(15)计算特征加权向量 $\boldsymbol{w}^t$,其中,$\boldsymbol{L}_s=\boldsymbol{D}-\boldsymbol{S}$;
4. 固定 $\boldsymbol{w}^t$,根据公式(20)更新相似矩阵 $\boldsymbol{S}^t$ 并计算投影矩阵 $\boldsymbol{F}^t$;
5. 如满足收敛条件或迭代 $t\geqslant I_m$,则输出结果,算法中止;反之,令 $t=t+1$,转至第 3 步.

3 LRAFL 算法描述

通过算法 1 可见,LRAFL 在实施过程中包含平衡参数γ,β,λ以及有效特征数 d 等待定参数,各类参数的优选过程不仅耗时而且对算法在不同数据集中的输出效果影响较大.因此,分析不同参数的具体实现推荐值是一个公知问题.此外,算法的收敛性和复杂度分析也对其具体的应用推广有着较大的影响.

3.1 参数设定细节

从公式(15)可见,特征加权向量 $\boldsymbol{w}$ 的取值由有效特征数 $d\in(0,m]$和正则项平衡参数$\gamma>0$ 决定.具体实施过程中,可根据输入数据对其中一个参数进行指示推荐,减少算法的计算开销.首先,当输入纯净数据时,可以认为所有的特征都是有效的,不同维数依 w_i 的取值具有不同的贡献度,即 $d=m$.不失一般性,假设 $w_1\geqslant w_2\geqslant\ldots\geqslant w_m\geqslant 0$ 按照从大到小的顺序排列,依特征加权的非负性,令 $w_m>0$,则有:

$$w_m=\frac{2\gamma m-z_d m+\sum z_j}{2\gamma m}\geqslant 0\Rightarrow\gamma\geqslant\frac{1}{2}\left(z_m-\sum z_j/m\right) \tag{21}$$

其中,在每次迭代中,$z_m-\sum z_j/m$ 的取值已知,且可以在 w_i 的计算过程中与其他部分进行抵消.因此,在此情形下,d 和γ的取值具有明确的指示,即 $d=m$ 且 $\gamma=\theta\left(z_d-\sum z_j/d\right)$,其中,$\theta\geqslant 1/2$ 是仅存的辅助变量.其次,当输入非纯净数据时,则认为 $\boldsymbol{w}$ 中部分特征的权值为 0,用于消除噪声干扰,即 $d<m$,则有:

$$w_{d+1}=\frac{2\gamma d-z_{d+1}d+\sum z_j}{2\gamma d}=0\Rightarrow\gamma=\frac{1}{2}\left(z_{d+1}-\sum z_j/d\right) \tag{22}$$

将其中的γ代入公式(15),得到 $\boldsymbol{w}$ 的最终计算方法为

$$\boldsymbol{w}=(-z+z_{d+1})\Big/\left(z_{d+1}-\sum_{j=1}^{d}z_j\right) \tag{23}$$

其中,仅存的人工设定参数 d 具有明确的物理意义,可依据输入数据按经验设定.

从公式(20)可见,相似矩阵中列向量 $\boldsymbol{s}_i$ 的取值由正则项参数β_i>0 决定.一般情况下,当得到的相似矩阵 $\boldsymbol{S}$ 全连通时,根据定理 1 可知数据为单簇结构,无法直接获得 $\boldsymbol{F}\in R^{n\times c}$ 矩阵.此外,在实际应用中,数据局部邻域关系更能刻画本质结构,往往仅考虑数据点 $\boldsymbol{x}_i$ 的 k 个邻域样本而非所有输入数据进行连接,而且稀疏的相似矩阵还能有效降低后续过程的计算量.因此,公式(20)中的 n 可由 k 替代且 $k<<n$.不失一般性,令 $\boldsymbol{s}_i$ 按降序排列,则有:

$$\left.\begin{matrix}s_{ik}>0\\ s_{i,k+1}\leqslant 0\end{matrix}\right\}\Rightarrow\left.\begin{matrix}\left(2\beta_i-kg_{ik}+\sum_{j=1}^{k}g_{ij}\right)\Big/2k\beta_i>0\\ \left(2\beta_i-kg_{i,k+1}+\sum_{j=1}^{k}g_{ij}\right)\Big/2k\beta_i\leqslant 0\end{matrix}\right\} \tag{24}$$

进一步可知,正则项平衡参数β_i的取值范围为

$$\frac{k}{2}g_{ik}-\frac{1}{2}\sum_{j=1}^{k}g_{ij}<\beta_i\leqslant\frac{k}{2}g_{i,k+1}-\frac{1}{2}\sum_{j=1}^{k}g_{ij} \tag{25}$$

因此,为获得最优解 $\boldsymbol{s}_i$,取:

$$\beta_i=\frac{k}{2}g_{i,k+1}-\frac{1}{2}\sum_{j=1}^{k}g_{ij} \tag{26}$$

再取β为$\beta_1,\beta_2,\ldots,\beta_n$的均值,即:

$$\beta=\frac{1}{n}\sum_{i=1}^{n}\beta_i=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}g_{i,k+1}-\frac{1}{2}\sum_{j=1}^{k}g_{ij}\right) \tag{27}$$

由于 k 是正整数并且有明确的物理意义,因此公式(20)仅需调整 k 求相似度矩阵,比直接调整β更为便捷.

从命题 1 可知,$Tr(\boldsymbol{F}^T\boldsymbol{L}_s\boldsymbol{F})$与 $rank(\boldsymbol{L}_s)=n-c$ 等价,因此在目标函数的更新过程中,取足够大的λ参数值时,$Tr(\boldsymbol{F}^T\boldsymbol{L}_s\boldsymbol{F})$无限接近于 0,可直接获得具有 c 分簇结构的相似矩阵 $\boldsymbol{S}$.因此,λ的取值可在算法运行中自适应确定,随机给定一个初始化值λ(如$\lambda=\beta$),每次迭代计算投影矩阵后,分别计算 $\rho_1=\sum_{i=1}^{c}\sigma_i(\boldsymbol{L}_s)$ 以及 $\rho_2=\sum_{i=1}^{c+1}\sigma_i(\boldsymbol{L}_s)$.给定接近于 0 的常数$\varepsilon$(本文选为 1e−10),当$\rho_1>\varepsilon$时,说明 $Tr(\boldsymbol{F}^T\boldsymbol{L}_s\boldsymbol{F})$值不够接近于 0,则增加$\lambda$值;反之,当$\rho_2<\varepsilon$时,说明 $Tr(\boldsymbol{F}^T\boldsymbol{L}_s\boldsymbol{F})$值过小,则减少$\lambda$值;当$\rho_1<\varepsilon<\rho_2$ 时,说明 $\boldsymbol{L}_s$ 矩阵恰好具有 c 块对角结构,模型收敛.

综上所述,虽然在算法 1 的描述中 LRAFL 有 4 个待设参数,但算法具体实施过程中仅 d(或γ)值和 k 值需要作调整测试,而且各参数都有明确的意义和设置推荐,保证算法应用过程的快速实现.

3.2 收敛性和复杂度分析

LRAFL 采用交替更新法进行模型迭代求解,在固定部分变量的前提下优化余下未知变量.根据算法 1 的描述,每次迭代的关键步骤公式(15)和公式(20)都是闭式解,因此其单个变量更新是唯一解.命题 2 说明所提算法在迭代过程中使目标函数(12)的值逐步下降,并最终收敛.

命题 2. 算法 1 的目标函数值随迭代过程逐步下降.

证明:假设在迭代 t 时有相似矩阵 $\boldsymbol{S}^t$,则在 t+1 次迭代中,固定 $\boldsymbol{S}^t$ 并优化 $\boldsymbol{F}^{t+1}$ 和 $\boldsymbol{w}^{t+1}$,以下不等式成立:

$$\begin{aligned}&Tr(\boldsymbol{W}^t\boldsymbol{X}\boldsymbol{L}_s^t\boldsymbol{X}^T)+\beta\boldsymbol{S}^t\odot\boldsymbol{S}^t+\gamma\|\boldsymbol{w}^t\|_2^2+\lambda Tr(\boldsymbol{F}^{tT}\boldsymbol{L}_s^t\boldsymbol{F}^t)\leqslant\\&Tr(\boldsymbol{W}^{t+1}\boldsymbol{X}\boldsymbol{L}_s^t\boldsymbol{X}^T)+\beta\boldsymbol{S}^t\odot\boldsymbol{S}^t+\gamma\|\boldsymbol{w}^{t+1}\|_2^2+\lambda Tr(\boldsymbol{F}^{(t+1)T}\boldsymbol{L}_s^t\boldsymbol{F}^{t+1})\end{aligned} \tag{28}$$

类似地,在固定 $\boldsymbol{F}^{t+1}$ 和 $\boldsymbol{w}^{t+1}$ 时优化相似矩阵,则有不等式:

$$\begin{aligned}&Tr(\boldsymbol{W}^{t+1}\boldsymbol{X}\boldsymbol{L}_s^t\boldsymbol{X}^T)+\beta\boldsymbol{S}^t\odot\boldsymbol{S}^t+\gamma\|\boldsymbol{w}^{t+1}\|_2^2+\lambda Tr(\boldsymbol{F}^{(t+1)T}\boldsymbol{L}_s^t\boldsymbol{F}^{t+1})\leqslant\\&Tr(\boldsymbol{W}^{t+1}\boldsymbol{X}\boldsymbol{L}_s^{t+1}\boldsymbol{X}^T)+\beta\boldsymbol{S}^{t+1}\odot\boldsymbol{S}^{t+1}+\gamma\|\boldsymbol{w}^{t+1}\|_2^2+\lambda Tr(\boldsymbol{F}^{(t+1)T}\boldsymbol{L}_s^{t+1}\boldsymbol{F}^{t+1})\end{aligned} \tag{29}$$

联合公式(28)和公式(29)可知,目标函数(12)的值随迭代过程逐步下降,命题 2 得证. □

值得注意的是:为避免 LRAFL 算法进入局部收敛,可以尝试不同的初始化方案,例如 $\boldsymbol{w}$ 可以简单地初始化为元素值为 $1/d$ 的列向量,亦可在非负加和约束下取随机值.此外,还可以在迭代循环外先初始化相似矩阵 $\boldsymbol{S}$,包括 k 近邻法或ε邻域法等,依不同输入数据集尝试不同的初始化方案,能使 LRAFL 有效逼近全局最优解.

算法 1 的关键耗时步骤是 3 个未知变量的更新操作,包括 $\boldsymbol{w}$,$\boldsymbol{S}$ 和 $\boldsymbol{F}$.其中,$\boldsymbol{w}$ 依公式(15)计算,其运算复杂度是 $O(d)$;$\boldsymbol{S}$ 中的列向量依公式(20)计算,其运算复杂度是 $O(k)$,因此相似矩阵 $\boldsymbol{S}$ 的整体复杂度是 $O(k^2)$;投影矩阵 $\boldsymbol{F}$ 通过 Laplacian 矩阵的特征分解获得,其复杂度是 $O(n^3)$.一般情况下,$d<<n^3$ 且 $k<<n$ 成立,因此 LRAFL 每次迭代的复杂度是 $O(n^3)$.假设算法的实际迭代次数是 I_m,可得 LRAFL 的综合运算复杂度是 $O(n^3I_m)$.

4 LRAFL 算法描述

4.1 合成数据实验

首先人工产生了 5 类独立的子空间,其环境维数为 250,本质维数为 4.对任意子空间,随机产生 100 个单位样本并将其中的 50%叠加高斯噪声干扰,噪声等级为{0,0.3,0.6}.图 1 显示了几种具有相似性矩阵构建能力的聚类算法所生成的邻域图,包括 LPS[13],RSS[25],LRS[24]和 LRAFL.

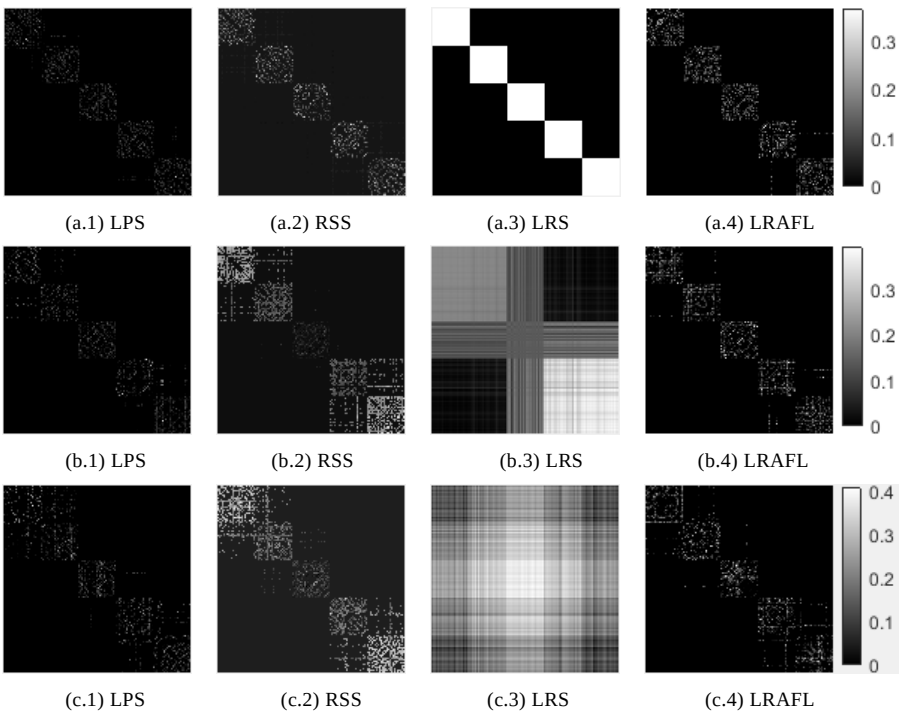


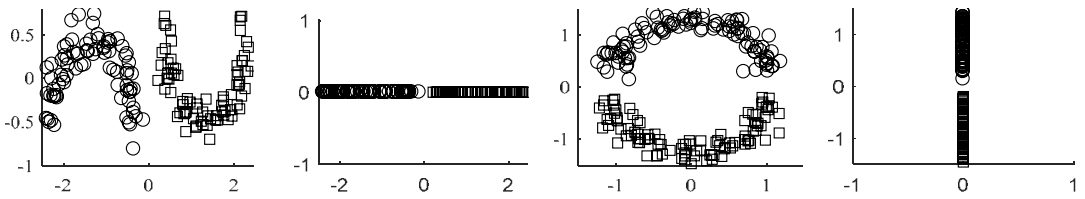
Fig.1 Affinity on synthesized data with different levels of noise corruption

图 1 五簇合成数据在不同噪声等级下的相似性结构

其中,图 1(a)~图 1(c)分别是无噪声干扰、30%噪声干扰和 60%噪声干扰下的效果,所有算法的参数优选过程遵从第 4.2 节的描述.从图 1 可见:在第 1 行无噪声干扰环境下,4 种算法都获得了高质量的相似度矩阵,为实现高性能的聚类结果奠定基础.然而,随着高斯噪声的引入, LRS 的相似矩阵完全处于紊乱状态,无法体现 5 分簇结构.类似地,LPS 的相似矩阵也趋于模糊,由 5 分簇结构逐渐退化为 3 分簇结构;RSS 的关联矩阵在趋于模糊的基础上,不同组结构的相似度值亦呈现不平衡特性.对比可见:所提算法 LRAFL 具有更为清晰的 5 簇相似度矩

阵,受噪声干扰的影响小于其他几种算法.值得注意的是:从图 1(a.3)可见,LRS 在干净环境下的相似矩阵非常清晰.然而,其类内相似度完全一致,说明 LRS 对同簇数据不具备多态区分性,解释了其较弱的抗噪能力.

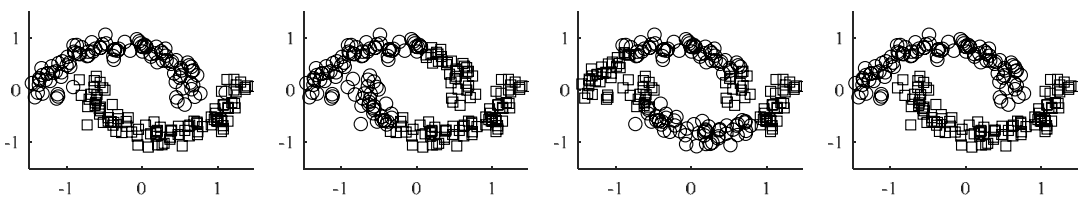
为进一步说明 LRAFL 的特征选择和数据聚类能力,采用人造的双半环数据进行效果验证.数据分为 2 簇,每簇 100 个样本点并随机叠加 15%的高斯白噪声.图 2 显示了 LRAFL 在不同位置分布情形下的双半环数据特征选择结果,其中,图 2(a)将数据左右放置,图 2(b)将其投影至高特征权值对应的坐标轴,图 2(c)是上下分布的数据,图 2(d)同理将其投影至高权值对应的坐标.可见:当输入数据分别处于左右和上下分布时,LRAFL 分别以横坐标和纵坐标作为高权值特征,说明其具有鉴别特征选择效果.此外,图 3 将双半环数据交叉放置,并采用 RSS, LRS 和 LRAFL 进行聚类对比,从中可见 RSS 和 LRS 两者的聚类结果都存在明显的错误,而 LRAFL 的聚类结果与输入簇结构完全吻合,表明其成功地将原始数据分成了 2 个类别,聚类效果优于 RSS 和 LRS.



(a) 左右半环 (b) 优选特征 (c) 上下半环 (d) 优选特征

Fig.2 Feature selection of LRAFL under different distribution of two-moon synthetic data

图 2 不同位置分布情形下的双半环数据特征选择结果



(a) Original (b) RSS (c) LRS (d) LRAFL

Fig.3 Clustering results on the two-moon synthetic data by RSS, LRS, and LRAFL

图 3 交叉分布情形下的双半环数据聚类效果对比

4.2 真实数据实验

通过 7 个不同的数据集和 11 种算法验证 LRAFL 模型的聚类性能,即准确度(AC)[25]和归一化互信息(NMI)[31]两个指标.测试数据包含 3 个人脸数据集(Orl,YaleB[23],Jaffe[12])、1 个语音字母数据集(Isolet)、2 个生物数据集(Yeast,Lung)和 1 个对象数据集(Coil[12]).为方便横向对比,所有带参考文献的数据集都依原文进行预处理,余下数据则保持原始形式.表 1 给出了各数据集的细节描述.对比算法包含 LLCFS[11],MCFS[8],NDFS[15], FSASL[12],LPS[13],kmeans[32],DEC[14],SMR[23],LRS[24],RSS[25]和 NSLLRR[22].

Table 1 Summary of the benchmark datasets and the number of selected features

表 1 数据集描述和实验中的特征选择数

数据集	样本数	特征数	类别数	特征选择
YaleB	640	2 016	10	[300,500,700,900,1100,1300,1500]
Jaffe	213	676	10	[70,140,210,280,350,420,490]
Isolet	1 559	617	26	[70,140,210,280,350,420,490]
Yeast	1 484	8	10	[1,2,3,4,5,6,7]
Lung	203	3 312	5	[300,600,900,1200,1500,1800,2100]
Coil	1 440	1 024	20	[120,240,360,480,600,720,840]
Orl	400	1 024	40	[120,240,360,480,600,720,840]

为获得各算法的最优实验结果,在实现时,需要对其人工参数进行网格搜索.在实验中,所有算法的正则参数和邻域参数范围分别设为{10^{-2},…,10^{2}}和{3,6,9,12,15}.表 2 和表 3 分别给出了所有对比算法通过 10 次随机初始化获得的准确度和归一化互信息指标.图 4 列出了各算法的平均 AC 和 NMI 指标.

Table 2 Aggregated clustering results measured by AC (%) of the competing methods

表 2 所有算法在不同数据集下的聚类准确度指标对比

Method	YaleB	Jaffe	Isolet	Yeast	Lung	Coil	Orl
k-means	15.28	71.57	49.64	35.08	78.33	59.17	51.79
LLCFS	20.47	77.93	52.02	37.20	*90.64*	60.49	51.75
MCFS	20.00	*78.87*	48.17	38.54	90.15	59.44	51.94
NDFS	19.56	75.80	56.96	37.20	**91.13**	64.65	53.58
FSASL	26.41	78.76	57.47	***53.44***	*90.64*	66.67	54.75
LPS	25.00	75.12	47.72	36.86	80.30	60.90	47.75
DEC	16.72	*95.31*	58.56	44.98	83.74	*73.06*	**62.25**
SMR	**62.19**	77.46	**58.63**	43.67	87.19	64.72	*61.75*
LRS	39.06	71.83	26.62	42.52	80.30	49.72	53.25
RSS	***65.78***	32.39	45.86	40.30	80.30	**76.46**	21.25
NSLLRR	62.30	**98.59**	***58.94***	*46.80*	*90.64*	62.01	***65.35***
LRAFL	*63.44*	***99.20***	*58.10*	**49.73**	***92.66***	***90.14***	61.50

Table 3 Aggregated clustering results measured by NMI (%) of the competing methods

表 3 所有算法在不同数据集下的聚类互信息指标对比

Method	YaleB	Jaffe	Isolet	Yeast	Lung	Coil	Orl
k-means	4.71	81.52	70.0	25.77	60.37	75.58	74.26
LLCFS	13.10	79.96	71.15	27.12	74.50	76.11	73.72
MCFS	13.80	83.62	69.55	**28.62**	73.35	76.21	74.93
NDFS	12.45	86.18	***75.78***	27.64	*75.39*	78.90	75.47
FSASL	21.74	*88.29*	**75.64**	27.32	74.76	79.20	75.88
LPS	21.48	84.93	70.60	27.64	60.54	75.98	71.56
DEC	8.67	*94.20*	73.48	27.22	68.25	*81.05*	*79.55*
SMR	57.31	83.34	72.92	27.37	69.80	76.60	*76.43*
LRS	34.53	78.33	38.98	22.05	58.24	56.66	75.12
RSS	***66.66***	28.52	64.58	27.33	54.27	**91.95**	41.76
NSLLRR	*57.60*	**97.81**	70.39	*28.40*	**76.12**	72.47	**79.86**
LRAFL	**64.50**	***98.76***	*74.81*	***36.08***	***80.82***	***97.05***	***80.23***

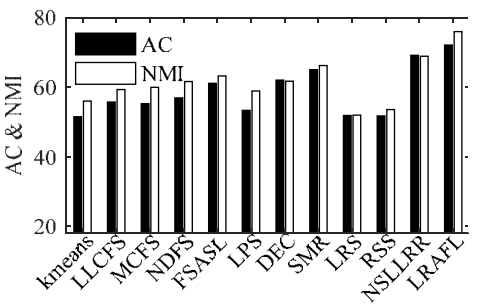


Fig.4 Average clustering results measured by AC and NMI (%) of the competing methods

图 4 各算法的平均聚类准确度和归一化互信息对比

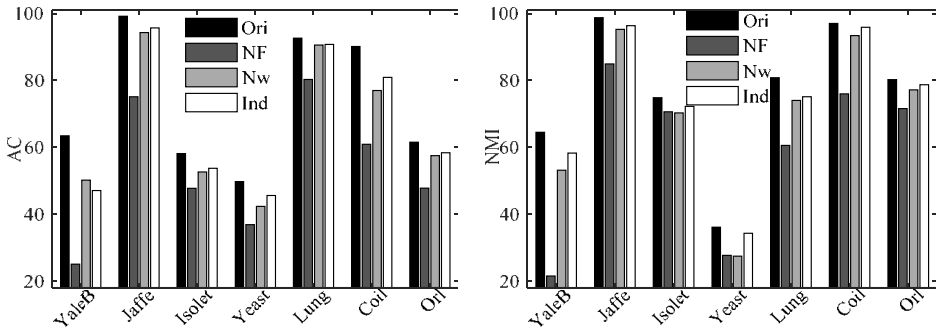
从表 2 和表 3 可知,

- 首先,依据样本的分布结构及特征量差异,各算法的聚类性能落差较大.例如:Yeast 数据集的维数较低,其类间间隔相对紧凑;而 YaleB,Orl 受光照、表情等影响较大.LRAFL 在这 3 个数据集中的 AC 指标分别仅为 49.73%,63.44%和 61.50%.此外,Lung 和 Jaffe 因为具有直观的分布结构比较容易实现聚类,其对应的 LRAFL 聚类 AC 指标分别达到了 92.66%和 99.20%;
- 其次,对比经典的 *k*-means 算法,DEC,NDFS 和 FSASL 等特征选择型算法都提升了聚类性能,而重构表示型算法,如 SMR 和 NSLLRR 则具有更加优秀的结果.RSS 受表示系数次优解影响,性能落差较大,在 YaleB 中获得了 65.78%的最高聚类精度,但在 Jaffe 和 Orl 中则仅获得了 32.39%和 21.25%的 AC 结果,

逊于基准模型 *k*-means;

- 最后,LRAFL 兼顾了特征优选机制和块对角 Laplacian 目标矩阵,其综合性能优于其他算法,在 AC 和 NMI 中分别赢得了 4 个和 6 个最高值,尤其在 Jaffe 和 Coil 中,分别获得了 98.76%和 97.05%的最高 NMI. 图 4 进一步表明,LRAFL 在各数据上的综合 AC 和 NMI 指标高于其他算法.

LRAFL 聚类模型包含特征优选权值 $\boldsymbol{w}$ 和簇结构逼近投影矩阵 $\boldsymbol{F}$ 用于块对角结构的相似度矩阵 $\boldsymbol{S}$ 构建.为进一步评估各子项的贡献度以及联合 $\boldsymbol{w}$,$\boldsymbol{S}$ 迭代更新的优势,将 LRAFL 算法分为无特征权值版(Nw)、无投影矩阵版(NF)、独立优化版 Ind 以及原始 LRAFL 模型(Ori),其中,Nw 是对目标函数式(12)中的特征加权部分去除,联合优化 $\boldsymbol{S}$ 和 $\boldsymbol{F}$;NF 指剔除公式(12)中的投影矩阵 $\boldsymbol{F}$,联合优化 $\boldsymbol{w}$ 和 $\boldsymbol{S}$.图 5 对比了 4 个版本在真实数据集中的准确度和归一化互信息指标.从图 5 可知,LRAFL 原始版在所有测试数据中的聚类指标都高于其减化版本.此外,NF 在各 LRAFL 减化版中效果最差,说明块对角 Laplacian 矩阵结构对聚类问题的关键性,与正文理论分析一致;Nw 版的聚类效果略逊于 Ind 版,说明特征优选机制能改善聚类分析效果;最后,Ind 版与 Ori 版的性能差距则进一步验证了 LRAFL 模型联合迭代更新的优势.

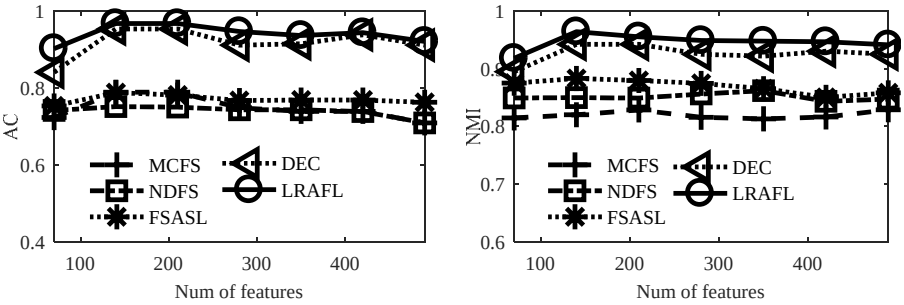


(a) AC (b) NMI

Fig.5 Clustering accuracy and NMI w.r.t. different versions of LRAFL

图 5 不同版本的 LRAFL 算法准确度和归一化互信息对比

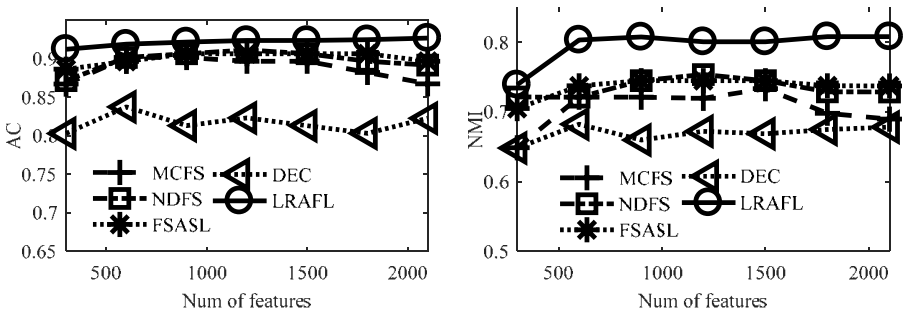
为进一步验证所提算法性能在不同特征选择量下的表现,图 6 将 LRAFL 与其他几种性能较优的特征选择算法进行对比,包括 MCFS,NDFS,FSASL 和 DEC,选用的数据集为 Jaffe 和 Lung,其中,前者的特征维数较少(676),后者的特征维数较高(3 312).从图 6 可见,LRAFL 在不同特征空间中的 AC 和 NMI 指标都优于其他算法,说明自适应特征学习机制能够有效地区分输入高维特征的优劣,进一步优化性能.随着有效特征维数的增加,不同算法的聚类性能都有所提升.然而,当维数进一步增加时,冗余特征导致算法的性能不增反降.相比较而言,LRAFL 除特征选择之外又添加了特征有效性加权机制,因此其精度曲线较为平坦.



(a) Clustering accuracy on Jaffe (b) Clustering NMI on Jaffe

Fig.6 Clustering accuracy and NMI w.r.t. different selected features

图 6 不同特征寻优下的分簇准确度和归一化互信息对比



(c) Clustering accuracy on Lung (d) Clustering NMI on Lung

Fig.6 Clustering accuracy and NMI w.r.t. different selected features (Continued)

图 6 不同特征寻优下的分簇准确度和归一化互信息对比(续)

4.3 算法效率分析

运行效率是算法应用能力的另一关键指标,本节选择在不同数据集中综合聚类性能表现较为突出的几种算法进行计算效率对比.指标测试时含参数优选过程,表 4 显示了各算法在各数据中的运行时间(单位:s).

算法运行平台为 Intel Core i5 CPU,双核主频 2.80GHz,内存 4GB,32 位 Win7 操作系统和 Matlab2014b 软件环境.

根据表 4 结果并结合表 2、表 3 可知:LRAFL 不仅在综合聚类效果上优于对比算法,而且其运行效率也具有明显的优越性.以 Jaffe 为例,LRAFL 仅需要 4.76s 完成参数优选和聚类分析运算,而排名第 2 的 NDFS 算法则耗费了近 80s 时间.此外,部分表示型聚类算法,如 SMR,RSS 和 NSLLRR,以运行效率为代价,在 YaleB, ISOLET 等数据集中取得了较 LRAFL 略优的聚类效果,但从表 4 可见,其运行时间呈指数级增长,基本较 LRFAL 慢 10 倍以上,尤其是 NSLLRR 模型,其综合聚类性能高于除 LRAFL 的其他对比算法,但是受非负系数矩阵构建以及多个人工可调参数影响,运行效率远远低于所有竞争算法,严重影响其应用扩展能力.

Table 4 Aggregated results measured by elapsed time of the competing methods

表 4 所有算法在不同数据集下的运行效率对比

Method	YaleB	Jaffe	Isolet	Yeast	Lung	Coil	Orl
MCFS	1.074e3	82.23	334.02	270.11	1.277e3	1.445e3	9.038e3
NDFS	1.146e3	78.18	549.91	252.59	3.649e3	1.017e3	240.59
FSASL	2.945e3	343.92	4.234e3	3.498e3	8.834e3	5.481e3	940.42
DEC	1.531e4	89.88	1.637e3	2.213e3	1.250e4	1.295e3	386.41
SMR	941.22	139.97	7.702e3	3.508e3	1.221e3	4.375e3	1.004e3
RSS	7.202e3	476.5	2.878e4	1.803e4	6.491e3	2.054e4	2.525e3
NSLLRR	2.896e5	2.768e4	2.567e6	8.256e5	6.432e4	2.168e6	1.105e5
LRAFL	155.34	4.76	17.17	170.96	242.81	191.78	24.19

4.4 参数敏感度分析

根据上述实验结果所示,所有聚类算法都有不同的人设参数待选,对算法应用效果和效率都有极大的影响.因此,所提算法的参数个数及其对不同设定值的敏感性是影响算法应用能力的又一指标.LRAFL 算法有两个待选参数——邻域数 k 和正则数γ,图 7 给出了其在不同选值范围下的聚类准确度性能变化,选用的数据集包括 YaleB,Jaffe,Yeast 和 Orl.

从图 7 结果可知:LRAFL 算法的两个参数中,γ对不同选值的敏感度较小,而且其选择过程也较为直观.本文采用 $2^x<300$ 取值,其中,$x\in\{-1,2,4,8,16\}$.k 对不同选值的敏感度较大,但具体应用过程中,k 的取值范围非常清晰,一般以 10 为中心向两边测试,减少了应用难度.此外,邻域数 k 是所有算法的待定参数,如何对其进行优选仍是一个公知问题.

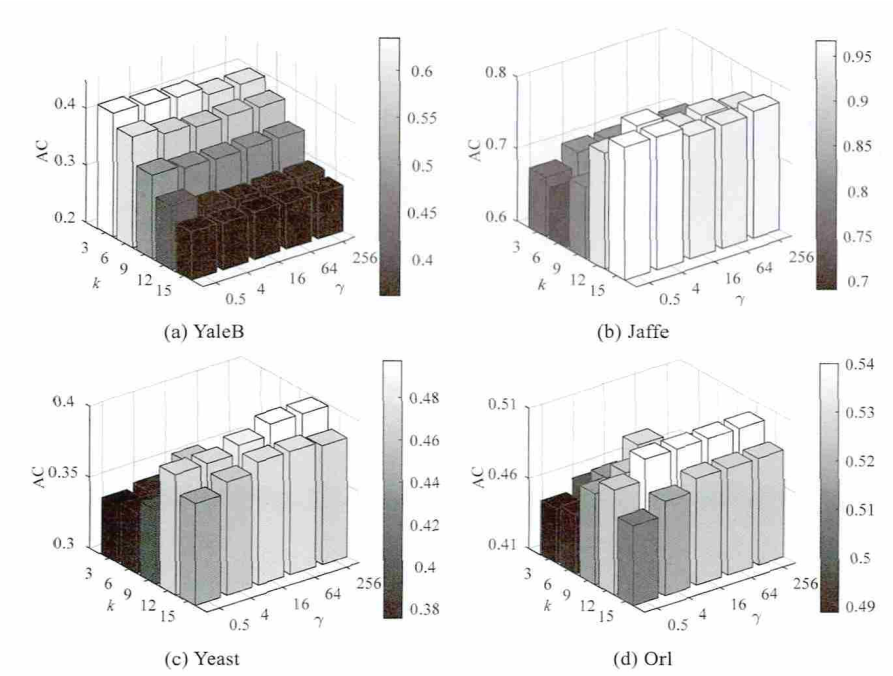


Fig.7 Clustering accuracy of LRAFL w.r.t. different parameters

图 7 LRAFL 参数优选下的聚类准确度

5 总 结

本文提出了一种新的数据聚类算法 LRAFL,兼顾自适应特征优选和簇结构学习两个关键目标.特征优选通过输入数据的重构表示进行自适应权值计算,并依权值高低进行有效特征筛选.簇结构学习通过对 Laplacian 矩阵强制进行 c 秩约束,获得精确的数据相似度矩阵,直接进行 c 簇结构划分.LRAFL 能够同时进行特征选择和数据聚类,且模型待设参数的物理意义明确,实现过程简洁直观.此外,设计了一种快速高效的模型求解算法,并给出了相应的算法复杂度分析和收敛性分析.通过大量人工合成数据和现实公开数据集验证了所提算法在精度、归一化互信息、运行效率和参数敏感度上较现存算法具有明显的优势.对比特征选择型算法,LRAFL 在聚类效果和运行效率上都具有优越的实验结果;对比表示型算法,LRAFL 虽然在部分数据集中无法获得更高的精度指标,但其运行效率却具有指数级的提升.

通过实验发现,所提算法 LRAFL 在应用过程中需要人工设定参数 k 和γ,虽然可以通过经验方式进行指引设置,且γ的取值对最终结果的影响较小,但仍然会削弱所提算法的应用扩展能力.因此,后续将集中进行待定参数的自适应确定或参数简化工作.此外,各聚类算法在不同的数据集中表现差异较大,不同先验样本分布对算法的性能影响仍不清楚,对其进行理论分析也是后续的工作之一.

References:

[1] Wu LF, He JY, Jian M, Zou YZ, Zhao TS. Local clustering analysis based FCN-CNN for cloud image segmentation. Ruan Jian Xue Bao/Journal of Software, 2018,29(4):1049−1059 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5409.htm [doi: 10.13328/j.cnki.jos.005409]

[2] Zheng JW, Zhu WB, Wang WL, Chen WJ. Smooth clustering with block-diagonal constrained laplacian regularizer. Journal of Computer-Aided Design & Computer Graphics, 2018,30(1):116−123 (in Chinese with English abstract).

[3] Li ZR. Research on generative adversarial network and its applications [Ph.D. Thesis]. Hangzhou: Zhejiang University of Technology, 2018 (in Chinese with English abstract)

[4] Tao H, Hou CP, Nie FP, Jiao YY. Effective discriminative feature selection with nontrivial solution. IEEE Trans. on Neural Networks and Learning Systems, 2016,27(4):796−808. [doi: 10.1109/TNNLS.2015.2424721]

[5] Yan H. Sparsity preserving score for joint feature selection. Applied Informatics, 2015, 2−8. [doi: 10.1186/s40535- 015-0009-3]

[6] Nie FP, Xiang SM, Jia YQ, Zhang CS, Yan SC. Trace ratio criterion for feature selection. In: Proc. of the 23rd National Conf. on Artificial Intelligence. AAAI Press, 2008. 671−676.

[7] Liu JW, Cui LP, Liu ZY, Luo XL. Survey on the regularized sparse models. Chinese Journal of Computers, 2015,38(7):1307−1325 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2015.01307]

[8] Cai D, Zhang CY, He XF. Unsupervised feature selection for multi-cluster data. In: Proc. of the 16th Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2010. 333−342. [doi: 10.1145/1835804.1835848]

[9] Shi CJ, Ruan QQ. Feature selection with enhanced sparsity for Web image annotation. Ruan Jian Xue Bao/Journal of Software, 2015,26(7):1800−1811 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4687.htm [doi: 10.13328/j.cnki.jos.004687]

[10] Hou CP, Nie FP, Li XL, Yi DY. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. IEEE Trans. on Cybernetics, 2014,44(6):793−804. [doi: 10.1109/TCYB.2013.2272642]

[11] Zeng H, Cheung Y. Feature selection and kernel learning for local learning-based clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(8):1532−1547. [doi: 10.1109/TPAMI.2010.215]

[12] Du L, Shen YD. Unsupervised feature selection with adaptive structure learning. In: Proc. of the 21th Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2015. 209−218. [doi: 10.1145/2783258.2783345]

[13] Yan H, Yang J. Locality preserving score for joint feature weights learning. Neural Networks, 2015,69:126−134. [doi: 10.1016/j.neunet.2015.06.001]

[14] Hou CP, Nie FP, Yi DY, Tao DC. Discriminative embedded clustering: A framework for grouping high-dimensional data. IEEE Trans. on Neural Networks and Learning Systems, 2015,26(6):1287−1299. [doi: 10.1109/TNNLS.2014.2337335]

[15] Li Z, Yang Y, Liu J, Zhou X, Lu H. Unsupervised feature selection using nonnegative spectral analysis. In: Proc. of the 26th Int'l Conf. on Artificial Intelligence. AAAI Press, 2012. 1026−1032.

[16] Wang F, Zhang CS. Label propagation through linear neighborhoods. EEE Trans. on Knowledge and Data Engineering, 2008,20(1):55−67. [doi: 10.1109/TKDE.2007.190672]

[17] Roweis ST, Saul LK. Noninear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326. [doi: 10.1126/science.290.5500.2323]

[18] Elhamifar E, Vidal R. Sparse subspace clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Press, 2009. 2790−2797. [doi: 10.1109/CVPR.2009.5206547]

[19] Liu GC, Lin ZC, Yan SC, Ma Y. Robust recovery of subspace structures by low-rank representation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,35(1):171−184. [doi: 10.1109/TPAMI.2012.88]

[20] Lu CY, Min H, Zhao ZQ, Zhu L, Huang DS, Yan S. Robust and efficient subspace segmentation via least squares regression. In: Proc. of the European Conf. on Computer Vision. IEEE Press, 2012. 347−360. [doi: 10.1007/978-3-642-33786-4_26]

[21] Feng JS, Lin ZC, Xu H, Yan SC. Robust subspace segmentation with block-diagonal prior. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Press, 2014. 3818−3825. [doi: 10.1109/CVPR.2014.482]

[22] Yin M, Gao JB, Lin ZC. Laplacian regularized low-rank representation and its applications. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016,38(3):504−517. [doi: 10.1109/TPAMI.2015.2462360]

[23] Hu H, Lin ZC, Feng JJ, Zhou J. Smooth representation clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Press, 2014. 3834−3841. [doi: 10.1109/CVPR.2014.484]

[24] Nie FP, Huang H. Subspace clustering via new low-rank model with discrete group structure constraint. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2016. 1874−1880.

[25] Guo XJ. Robust subspace segmentation by simultaneously learning data representations and their affinity matrix. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2015. 3547−3553.

[26] Zelnik-Manor L, Perona P. Self-Tuning spectral clustering. In: Proc. of the Advances in Neural Information Processing Systems 17. MIT Presss, 2004. 1601−1608.

[27] Cheng B, Yang JC, Yan SC, Fu Y, Huang TS. Learning with-graph for image analysis. IEEE Trans. on Image Processing, 2010, 19(4):858−866. [doi: 10.1109/TIP.2009.2038764]

[28] Huang J, Nie FP, Huang H. A new simplex sparse learning model to measure data similarity for clustering. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence. AAAI Presss, 2015. 3569−3575.

[29] Hagen L, Kahng AB. New spectral methods for ratio cut partitioning and clustering. IEEE Trans. on Computer-aided Design of Integrated Circuits and Systems, 1992,11(9):1074−1085. [doi: 10.1109/43.159993]

[30] Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8): 888−905. [doi: 10.1109/34.868688]

[31] Nie FP, Wang XQ, Huang H. Clustering and projected clustering with adaptive neighbors. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Presss, 2014. 977−986. [doi: 10.1145/2623330.2623726]

[32] Von Luxburg U. A tutorial on spectral clustering. Statistics and Computing, 2007,17(4):395−416.

[33] Cai S, Zhang L, Zuo W, Feng X. A probabilistic collaborative representation based approach for pattern classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Press, 2016. 2950−2959.

[34] Fan K. On a theorem of WEYL concerning eigenvalues of linear transformations II. Proc. of the National Academy of Sciences of the United States of America, 1950,36(1):31−35.

[35] Nocedal J, Wright S. Numerical Optimization. 2nd ed., New York: Springer-Verlag, 2006. [doi: 10.1007/978-0-387-40065-5]

附中文参考文献:

[1] 毋立芳,贺娇瑜,简萌,邹蕴真,赵铁松.局部聚类分析的 FCN-CNN 云图分割方法.软件学报,2018,29(4):1049−1059. http://www.jos.org.cn/1000-9825/5409.htm [doi: 10.13328/j.cnki.jos.005409]

[2] 郑建炜,朱文博,王万良,陈婉君.块对角拉普拉斯约束的平滑聚类算法.计算机辅助设计与图形学学报,2018,30(1):116−123.

[3] 李卓蓉.生成式对抗网络研究及其应用[博士学位论文].杭州:浙江工业大学,2018.

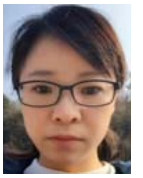
[7] 刘建伟,崔立鹏,刘泽宇,罗雄麟.正则化稀疏模型.计算机学报,2015,38(7):1307−1325. [doi: 10.11897/SP.J.1016.2015.01307]

[9] 史彩娟,阮秋琦.基于增强稀疏性特征选择的网络图像标注.软件学报,2015,26(7):1800−1811. http://www.jos.org.cn/1000-9825/4687.htm [doi: 10.13328/j.cnki.jos.004687]

郑建炜(1982－),男,浙江嵊州人,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘,模式识别,机器学习,数值最优化.

王万良(1957－),男,博士,教授,博士生导师,主要研究领域为智能科学,人工智能,大数据分析.

李卓蓉(1986－),女,博士,讲师,CCF 专业会员,主要研究领域为人工智能,大数据分析,深度学习.



陈婉君(1982－),女,讲师,主要研究领域为智能科学,数据分析.